"What if...": The use of conceptual simulations in scientific reasoning

Susan Bell Trickett, Naval Research Laboratory

Tel: 703 577 6035

Fax: 703 993 1300

J. Gregory Trafton, Naval Research Laboratory

Tel: 202 767 3439

Fax: 202 404 4080

Suggested running head: Conceptual Simulations

Keywords: scientific reasoning, scientific discovery, visualization, model-based reasoning, analogy, problem-solving, "in vivo" observation

Author Notes

Susan Bell Trickett is a postdoctoral fellow at the Naval Research Laboratory.

J. Gregory Trafton is a cognitive scientist at the Naval Research Laboratory.

Correspondence concerning this article should be addressed to Susan Trickett, NRL, Code 5515, Washington, DC 20375-5337 (trickett@itd.nrl.navy.mil)

Abstract

We use the term conceptual simulation to refer to a type of everyday reasoning strategy commonly called “what-if” reasoning. It has been suggested in a number of contexts that this type of reasoning plays an important role in scientific discovery; however, little *direct* evidence exists to support this claim. We propose that conceptual simulation is likely to be used in situations of informational uncertainty, and may be used to help scientists resolve that uncertainty. We conducted two studies to investigate the relationship between conceptual simulation and informational uncertainty. Study 1 was an “in vivo” study of expert scientists; the results suggest that scientists do use conceptual simulation in situations of informational uncertainty, and that they use conceptual simulation to make inferences from their data, using the analogical reasoning process of alignment by similarity detection. Study 2 experimentally manipulated experts’ level of uncertainty and provides further support for the hypothesis that conceptual simulation is more likely to be used in situations of informational uncertainty. Finally, we discuss the relationship between conceptual simulation and other types of reasoning using qualitative mental models.

1. Introduction

In a famous anecdote, Einstein describes how, as a youth, he visualized himself chasing a beam of light, and he explains that later on, this imaginative leap contributed to his development of the theory of relativity (Einstein, 1979). Einstein's thought experiment is one of the best-known examples of a type of "what if" reasoning that has been implicated in scientific discovery in a variety of fields. Other famous scientists who are reported to have engaged in thought experiments include Galileo, Newton, Maxwell, Heisenberg, and Schrödinger, to name a few (e.g. Shepard, 1988).

Scientists are likely to use such thought experiments, or "what if" reasoning, when it is either impossible or impractical to conduct a physical experiment. In addition, from a purely theoretical perspective, "what if" reasoning offers several advantages. Unlike quantitative reasoning strategies, it does not require numerical precision. This may be useful a) when precise quantitative information is not available or b) when a scientist is attempting to develop a general, or high-level, understanding of a system. Like other forms of mental-model-based qualitative reasoning, "what-if" reasoning allows one to reason with partial knowledge (whether incomplete or imprecise) and hence to accommodate the ambiguity inherent in situations of uncertainty (Forbus, 2002). "What if" reasoning also allows the construction of multiple alternatives, which may be useful in generating predictions or explanations when scientists lack principled knowledge that can allow them to proceed in their reasoning with some measure of certainty. All these situations share a high level of uncertainty; thus, "what if" reasoning may be especially useful in some situations of uncertainty.

There are many types of uncertainty in complex domains such as scientific enquiry (Schunn, Kirschenbaum, & Trafton, under review). Schunn et al. differentiate between subjective uncertainty (what a person feels) and objective uncertainty (uncertainty in the information a person has). Our focus here is on informational uncertainty.

For reasons discussed below, we concentrate our research on the data analysis phase of scientific discovery. During this phase, scientists must first recognize what information the data actually represent, and second, come to an understanding of what that representation actually means in terms of their research questions (i.e., interpret the data). Consequently, there are two general areas where scientists are likely to encounter informational uncertainty. First, the data themselves may literally be unclear: data may be missing, inaccurate, or noisy, for example, so that scientists must work to differentiate real phenomena from noise. Second, the *meaning* of the data may be unclear; for example, experimental results may be anomalous (i.e., incompatible with previously established empirical results or even theory), follow some unexpected or unusual pattern, or otherwise conflict with the scientist's predictions. Part of the scientist's task is to explain or otherwise resolve such expectation violations.

In other complex domains, such as meteorology, we have found that when people experience informational uncertainty when using complex visualizations, they mentally transform the visualization by adding their own representation of uncertainty, in order to resolve it (Trickett, Trafton, Saner, & Schunn, in press). Consequently, we expect that when scientists experience informational uncertainty, they will try to resolve that uncertainty, and we propose that “what if” reasoning is likely to be one strategy by which they attempt to do so, because it allows people to transform their current understanding by mentally constructing an alternative. “What if” reasoning allows people to think through the implications of different starting assumptions by playing out different scenarios and then to evaluate their plausibility. If this were the case, we would expect “what if” reasoning to occur particularly in association with tentative explanations, or hypotheses, that could account for particular instances of informational uncertainty. Furthermore, if “what if” reasoning were used to try to resolve such uncertainty, we would expect it to lead to an evaluation of the hypothesis, in order to determine whether it adequately accounts for the uncertainty and consequently resolves it.

What constitutes “what if” thinking? Brown (2002) proposes a three-step process that consists of first, visualizing some situation, second, carrying out one or more operations on it, and third, seeing what happens. The third part of the process—“seeing what happens”—is crucial. It distinguishes “what if” thinking from purely imagining, because during this third phase *causal reasoning* occurs to the results of the manipulation(s) of the second phase. A well-known example of this type of thinking is Lucretius’ attempt to show that space is infinite (Brown, 2002). Assuming space has a boundary (*visualize a situation*), throw an imaginary spear toward it (*carry out an operation on the visualization*). If the spear goes through, there is no boundary; if the spear rebounds, we infer a “wall” which must itself be in space that stopped the spear (*see what happens*). Consequently, space has no boundary (*causal reasoning*).

Although Lucretius is clearly not a layperson, it is easy to apply the same processes to everyday examples of this type of thinking. For example, suppose one is figuring out the steps by which to assemble a piece of furniture (e.g., Lozano & Tversky, 2006), in the absence of clear written instructions, and prior to making any irreversible decisions. One might mentally start to arrange certain pieces where one thinks they should go (*visualize a situation*). Then one might mentally attempt to insert a new piece (*carry out an operation on the visualization*). One can then inspect the visualization to determine whether the new piece will fit (*see what happens).* Finally, one can determine whether the initial arrangement is correct and decide either to proceed with construction or to start over (*causal reasoning)*.

As this illustration shows, "what if" thinking is hardly the type of arcane activity frequently associated in the popular imagination with scientific genius, but rather an everyday reasoning strategy available to scientist and layperson alike. How important is such a strategy likely to be in the scientific reasoning process? On the one hand, scientific expertise—domain knowledge and skills—is acquired only after many years of education and practice (Ericsson & Charness, 1994; Ericsson, Krampe, & Tesch-Roemer., 1993; Schunn and Anderson, 1999). On the other hand, current research suggests that, as Einstein himself maintained, what sets scientific reasoning apart from everyday reasoning is not different processes, but simply greater precision, systematicity, and "logical economy" (Klahr & Simon, 1999). A full model of scientific discovery should therefore include relevant everyday reasoning strategies and heuristics. It has already been suggested that everyday reasoning strategies, such as mental simulation and other forms of reasoning with qualitative mental models, play a role in a general understanding of natural phenomena and physical systems (e.g., Hayes, 1988; Williams & de Kleer, 1991). Our question is the extent to which one such strategy—"what if" reasoning—guides the reasoning of experts' scientific reasoning.

In fact, several everyday reasoning strategies have already been shown to play an important role in the process of science, strategies such as analogy (Dunbar, 1995, 1997; Gentner, 2002; Okada & Simon, 1998), attending to anomalies (Kulkarni & Simon, 1988), collaboration (Azmitia & Crowley, 2001), use of mental models (Forbus, 1983; Forbus & Gentner, 1997), and the like. The goal of this paper is to investigate the role of "what if" thinking in the scientific reasoning of contemporary scientists.

There is some evidence in the cognitive science literature that scientists specifically use forms of "what if" reasoning. Reconstructions of historical discoveries and analyses of contemporary records such as journals and lab notebooks suggest that scientists conduct "mental experiments" in a process that mirrors an empirical experiment (Nersessian, 1999) or otherwise construct "runnable" mental models (e.g., Ippolito & Tweney, 1995). Empirical studies of contemporary scientists also find the use of mental experiments (e.g.,Clement, 2002a; Qin & Simon, 1990) and mental simulation (Schraagen, 1993). This research spans a wide variety of contexts (such as historical reconstruction, protocol study, and lab experiment), tasks (such as scientific discovery, experimental design, and prediction), and participants (from famous historical figures to contemporary expert practitioners to scientists-in-training).

Despite this body of research, it is difficult to draw general conclusions from the results. The nature of historical studies makes it impossible to determine whether the mental experimentation occurred in the course of the problem-solving or retrospectively (Saner & Schunn, 1999). Nor are the studies of contemporary scientists conclusive. Qin and Simon told participants to generate a mental image prior to performing the task, so that their use of mental experimentation may not have been spontaneous. The scientists observed by Schraagen and by Clement were not experts in the specific task domain, and therefore lacked precise domain knowledge. The use of "what if" reasoning in these studies was clearly spontaneous; however, perhaps the scientists were using it to compensate for their lack of domain knowledge (i.e., in this case, conceptual simulation was more of a lay strategy than an expert one).

In sum, no experimental studies have been conducted with the express purpose of investigating the use of "what if" reasoning among expert, practicing scientists working in their own domain, and as a result, no clear picture has emerged as to when, how, and why scientists might use this strategy. Our goal is first, to gather evidence that expert scientists do, in fact, engage spontaneously in "what if" reasoning, and second, to investigate how they do so and how significant a role this strategy plays in their acts of scientific enquiry.

Researchers use many different terms to describe the strategy we have loosely discussed as "what if" reasoning—*mental experiment, thought experiment, inceptions, mental simulation*, and so on. In all cases, however, the underlying strategy demonstrates the characteristics described by Brown (2002), discussed above. In our study, we shall refer to these separate processes—visualizing a situation, carrying out mental operations on it, and seeing what happens—collectively as "conceptual simulation." We believe this term captures the two most crucial aspects of this type of reasoning, namely, it occurs at the conceptual level (rather than, say, in any actual or external sense), and it involves mentally playing out, or "running," a model of the visualized situation, in order that changes can be inspected.

More specifically, conceptual simulation involves constructing and manipulating a mental model that not only derives from an external representation but is also an analog of it (Clement, 2002b; Nersessian, 1999; Schwartz & Black, 1996a). Functionally, conceptual simulations adapt the external representation by adopting hypothetical values and playing out their implications, to move beyond the information actually represented. This process allows new inferences about that information to be made.

Our first challenge has been to develop a reliable means of identifying conceptual simulations, which are an internal cognitive process rather than a directly observable behavior. Our general method has been to collect verbal protocols of scientists solving problems in their own domain. This method is based on the assumption that contents of working memory are "dumped" into the speech stream, where they can be examined and coded (Ericsson & Simon, 1993). In order to increase the reliability of this detection and coding process, we have operationalized the notion of conceptual simulation such that the construct is empirically grounded and observable in the speech stream: in a continuous sequence of utterances, the scientist a) refers to a new representation of a system or mechanism, b) refers to transforming that representation spatially, in a hypothetical manner, and c) refers to a result of the transformation. This three-stage process corresponds to the processes described by Brown (2002) in defining “what if” thinking.

Our first study is exploratory, and examines the question of whether and to what extent practicing scientists spontaneously use conceptual simulations. We further investigate the extent to which this strategy is used to resolve specific instances of informational uncertainty, in a cycle of hypothesis statement and evaluation. In order to determine the significance of the relationship between “what if” reasoning and hypothesis evaluation, we investigate the frequency of use for other hypothesis-evaluation strategies that have been identified in the scientific reasoning literature. If “what if” reasoning plays a significant role in the hypothesis evaluation process, it should occur at least as frequently as these known strategies. Furthermore, there may be relationships between “what if” reasoning and these other strategies that can illuminate the overall process of resolving informational uncertainty. To foreshadow the outcome of Study 1, our results suggest that scientists do spontaneously use conceptual simulations, and they seem to do so as a means of resolving informational uncertainty. Study 2 is a laboratory experiment—also of expert scientists—in which we manipulate uncertainty in order to further test this hypothesis.

2. Study 1

2. 1. Method

Dunbar has demonstrated the value of naturalistic observation of scientists in uncovering previously underspecified strategies and dynamics in the science laboratory. We have therefore adapted Dunbar's "in vivo" methodology for on-line observation of scientific thinking, in which participants perform their regular tasks and the experimenter observes and records their interactions (Dunbar 1995, 1997). We have focused our investigation on one specific scientific task—data analysis—because it is a crucial task for many scientific domains, one during which scientists attempt to account for their data and in which they are likely to experience a great deal of informational uncertainty. Data analysis is therefore likely to produce a rich record of scientific thinking and hypothesis-generation about informational uncertainty.

2.1.1. *Participants*

Participants were recruited through personal connection of the experimenter or her associates. The sample of scientists was selected to represent a diverse array of fields, rather than just one particular subfield, and several different stages of data analysis, in order to make the results more generalizable. Observations were recorded from nine scientists in eight data analysis sessions. All the participants were either expert scientists who had earned their PhDs more than six years previously, or graduate students working alongside one of these experts. Four of the sessions involved an expert scientist working alone. Three of the group sessions involved a senior researcher and one or more graduate students; the remaining group session involved two expert scientists. (Some scientists were thus observed over more than one session.) Four sessions were in branches of physics (astronomy and computational fluid dynamics, or CFD), two were in neuroscience (fMRI and neural spikes), and two were in cognitive psychology. Of the three datasets pertaining to computational fluid dynamics, one focused on a problem involving a submarine and two focused on laser pellet research.

2.1.2. *Procedure*

Participants agreed to contact a member of the research team when they were ready to conduct some analysis of recently acquired data, and an experimenter visited the scientists at their regular work location. All participants agreed to be videotaped during the session. Participants working alone were trained to give talk-aloud verbal protocols. For scientists working in groups, their conversation was recorded as they engaged in scientific discussion about their data. All participants were instructed to carry out their work as though no camera were present and without explanation to the experimenter (Ericsson & Simon, 1993).

Details about each individual session are reported in Table 1. All utterances were transcribed and segmented according to complete thought (off-task utterances were excluded from analysis). Finally, a coding scheme (described below) was developed to explore the relationship between conceptual simulation, uncertainty, and hypothesis evaluation.

--------------------Insert Table 1 about here--------------------

2.1.3. *Analysis Tools and Tasks*

The psychology data were displayed numerically in Excel; all the other data were displayed using visualization tools specific to the domain. Fig 1. shows an example of the visualization software used by one of the physicists.

--------------------Insert Fig. 1 about here--------------------

Although each scientist or group of scientists used different tools, their tasks shared several characteristics. They were all analyzing data that they themselves had collected, from observations, from a controlled experiment, or from running a computational model. They displayed the data using their regular tools. Apart from the second CFD laser session, which was a follow-up to the first session, all sessions represented the initial investigation of these data. Whether their interest was exploratory or confirmatory, their goal was to understand the fundamental processes that underlay the data. Table 2 summarizes the characteristics of each data analysis session.

--------------------Insert Table 2 about here--------------------

2.1.4. *Coding Scheme*

The overall goal of this research was to investigate whether and when scientists use conceptual simulation, whether they use it to resolve informational uncertainty, and to what extent they do so, relative to other strategies. We predicted that scientists would use conceptual simulation to evaluate hypotheses they proposed to account for informational uncertainty. We therefore developed a coding scheme that would allow us to identify conceptual simulations, hypotheses, and several strategies that have been shown to be associated with hypothesis evaluation.

Conceptual simulation

A conceptual simulation spans several utterances. It begins with a reference to a representation of a system or part of a system. Mental operations are then carried out on this representation in order to simulate the system's hypothetical behavior under certain circumstances. The initial representation may be grounded internally (for example, in domain knowledge or memory of a previously observed phenomenon) or externally (for example, in a displayed image). However, simply forming and transforming a mental image is not sufficient. The key feature of a conceptual simulation is that it involves a simulation "run" that is both hypothetical (i.e., it does not merely reproduce observed behavior) and alters the starting representation, producing a different end state that can be inspected, in order to "see what happens" (cf Brown, 2002).

In order to formally code conceptual simulations, we adapted Trafton's spatial transformation framework (Trafton, Trickett, & Mintz, 2005). Spatial transformations occur when a spatial object is transformed from one mental state or location into another mental state or location. They occur in a mental representation that is an analog of physical space. They can be performed purely mentally (e.g., purely within spatial working memory or a mental image) or "on top of" an existing visualization (e.g., a computer-generated image). (See Trafton, Trickett, Stitzlein, Saner, Schunn, &

Kirschenbaum, 2006 for more on spatial transformations.) This initial representation provides the starting point for a conceptual simulation. Therefore, we first identified references to a new representation. We then performed a spatial transformation analysis on the utterances that immediately followed, in order to determine whether any mental operations were applied to transform that representation. Some possibilities include rotation, modification (by addition or deletion), moving an image, animating features, and comparison. Finally, we identified the reference to the result of the transformation(s). Conceptual simulations may thus be defined formally as a specific sequence:

1. refers to a new representation of a system or mechanism
2. refers to transforming that representation spatially, in a hypothetical manner
3. refers to a result of the transformation (seeing what happens).

Table 3 illustrates examples of conceptual simulation. Note that although a conceptual simulation spans several utterances, collectively these are coded as only one conceptual simulation. (See Table 4 for additional examples of conceptual simulation, at http://www.cognitivesciencesociety.org/supplements/.)

--------------------Insert Table 3 about here--------------------

Hypotheses

Every utterance was examined, and all statements that attempted to explain or account for a phenomenon were coded as hypotheses, for example, *"OK, so now he's not showing activation for the motor preparation,* ***so maybe that's just a function of it being the first thing he did****"* (source: fMRI; hypothesis in bold type).

Scientific reasoning strategies

We selected several strategies from the scientific reasoning literature: data focus, empirical test, consult a colleague, tie-in with theory/domain knowledge, and analogy. Data focused strategies are highly relevant to scientific inquiry in general and to data analysis in particular. Testing a hypothesis by empirical means is part of the scientific method (Popper, 1956) and has been much studied in the scientific reasoning literature (e.g., Klahr & Dunbar, 1988; Klahr & Fay, 1990; Schunn & Anderson, 1999; Vollmeyer Burns, & Holyoak, 1996). Collaboration (consulting a colleague) has been shown to be instrumental in solving scientific problems in both instructional and professional settings (Azmitia & Crowley, 2001; Okada & Simon, 1997). Domain knowledge is also an important factor in expert performance among scientists (Chinn & Malhhotra, 2001; Schunn & Anderson, 1999), as is a deep understanding of the tools, instruments, and techniques used in a given domain (Schraagen, 1993; Schunn & Anderson, 1999). Finally, research has identified analogy as a powerful reasoning mechanism for science (Dunbar, 1997; Forbus & Gentner, 1997; Nersessian, 1992a; Thagard, 1992).

Analogical reasoning involves mapping information from one domain or instance —the “source”—to another—the “target”—in order to make inferences about the target (Gentner, 1989). Different theories of analogy specify different processes by which the mapping between source and target occurs, for example, structural alignment (Gentner, 1983; Holyoak, 1985) constraint satisfaction (Holyoak & Thagard, 1989), and similarity detection (e.g., Gentner & Markman, 1997). During the mapping or alignment phase, regardless of the specific mechanism by which it occurs, the relevant parts of the source are "applied" to the target, and inferences about the target are drawn. Alignment thus involves an explicit or implicit comparison between two representations, and the detection of similarities between them.

Gentner & Markman (1997) propose that analogy and similarity are related through the process of structural alignment. The difference lies in the relative importance of relational similarity (in analogy) and attribute similarity (in similarity judgments). Whereas analogical comparisons focus primarily on structural or relational similarity, similarity judgments focus more on commonalities between attributes, or surface features. (Note that "mere-appearance matches" have no relations in common, and are therefore are not discussed further here.) Because of the visual/spatial nature of much of the data in these scientific domains, we expect the scientists to make a significant number of similarity judgments, in addition to more structurally focused analogical comparisons. According to Gentner and Markman, structural alignment guides the comparison process in both cases, analogy and similarity. Also, in both cases, the comparison process focuses on alignable differences, which allow a person to identify on relevant differences between the two entities being compared. We use the term "analogy" to refer to comparisons based primarily on structural or relational similarity, and the term "alignment" or "alignment by similarity detection" to capture the process of comparison based primarily on attribute similarity, in which one representation is matched up to another, in order to detect relevant areas of similarity and difference.

To code all these strategies, we identified all utterances that immediately followed a hypothesis that further elaborated the hypothesis, whether they supported or opposed it. Those utterances were coded as follows:

Data focus

Following Trafton et al. (2005), we coded statements that "read off" data from the visible display as data focus. Utterances that referred to looking at data in a different way (such as re-plotting the data or displaying it in a different visualization), to "tweaking" data (for example, by transformation or removing outliers, etc.), or to looking at data that were not currently on view but that were available were also coded as data focus strategies. See Table 5 for examples of data focus strategies.

--------------------Insert Table 5 about here--------------------

Empirical test

Utterances in which the scientist proposed to collect additional data were coded as empirical test strategies. These included experiment proposals, making plans to run a new experiment, planning to collect additional data for an existing experiment (for example, increasing the sample size), or planning to collect more observational data. Plans to build and run computational models were also coded as empirical test strategies. Table 6 illustrates the coding of empirical test strategies.

--------------------Insert Table 6 about here--------------------

Consult a colleague

Utterances that refer to showing the data to or asking the opinion of a co-worker or other expert were coded as consulting a colleague, for example, *“I’m gonna have to discuss it with, ah, John when he gets back. And with Bob”* (source: CFD—submarine). (Names have been changed to safeguard participants’ anonymity.)

Tie-in with theory/Domain knowledge

Utterances that referred to theoretical underpinnings of the data were identified and coded as tie-in with theory, for example, *"But just in general, if you have, I mean in your, your theoretical ring galaxy of the computer…"* (source: Astronomy). In addition, utterances that drew on domain-specific skills, such as an understanding of tools and techniques, were also included in this category, for example, *"Ah, I'm beginning to wonder if we didn't have enough velocity range all of a sudden"* (source: Astronomy).

Analogical reasoning

Analogical reasoning was coded using the definition and coding scheme developed by Dunbar (1997). According to this scheme, an analogy exists when a scientist either refers to another base of knowledge to explain a concept or uses another base of knowledge to modify a concept. Analogies were coded at a general level, when both source and target were explicitly identified (e.g., "The atom is like the solar system") and at the level of the alignment by similarity detection. Table 7 illustrates this coding of analogy.

--------------------Insert Table 7 about here--------------------

2.2. Results and Discussion

Eight datasets were analyzed, comprising 331 minutes of relevant protocol, broken into 3278 on-task utterances.

2.2.1. *Inter-rater reliability*

We used two approaches to establish inter-rater reliability. First, after one coder had coded *all* the data for conceptual simulations, a second, independent coder coded ten percent of the entire dataset, pinpointing any conceptual simulations. (The data to be coded were selected from two domains by the first coder, because they contained examples of conceptual simulations and of sequences that it might be challenging to determine whether they were conceptual simulations or not.) To illustrate this approach in the CFD domain, consider the set of utterances in Table 8. The first coder identified that lines 6 to 12 contained a conceptual simulation, in which the speaker was trying to reconstruct how one of the modes could have grown at a slower rate than another. The first coder ended the conceptual simulation at line 12, noting that in lines 13 and 14, the scientist aligned the end result of the mode being fed with the displayed representation of the final growth of the other modes involved. The remainder of this section was not coded as conceptual simulation, because the scientist is recalling theoretical information about the way the modes interact. The second coder then reviewed this entire section, embedded within a much larger context of several previous and subsequent utterances, to determine whether a conceptual simulation occurred, and if so, which utterances it spanned. We initially took this coarse-grained approach, in order to establish that conceptual simulations could be reliably isolated in the speech stream. This approach resulted in 98% agreement, $k = .91$, $p < .01$.

--------------------Insert Table 8 about here--------------------

Second, we performed a finer-grained analysis, coding for each utterance whether it was part of a conceptual simulation or not. The same first coder's ratings were used. We then selected thirty-three percent of the entire dataset for coding by yet a third independent coder. We divided each session's data into three equal parts, based on the number of utterances, and selected the first, second, or third section at random from each dataset. As a result, the third coder coded one-third of each dataset, and collectively, the sections represented early, middle, and late analysis on the part of the scientist. The first coder's ratings were not available to the third coder at any time during the process. To summarize the difference between the two rounds of coding, in the first, coarser-grained round, the second coder identified given *sequences* of utterances as comprising a conceptual simulation or not. In the second round, the third coder identified line by line whether or not each *utterance* was part of a conceptual simulation.

The third coder was trained to recognize conceptual simulations, by using examples that were not part of the to-be-coded data (see Appendix A for more information about the training). The coder examined each utterance and judged whether the speaker referred to a new representation, whether, immediately afterwards, the speaker referred to one or more mental operations that transformed that representation (spatial transformations), and whether the speaker referred to the result of those transformations. If the coder observed this sequence, the individual utterances were scored as part of a conceptual simulation. Utterances that did not contribute to this sequence were scored as "no conceptual simulation." The third coder worked entirely independently of the first coder. The coders conferred once after the third coder had coded one dataset, in order to resolve any questions or difficulties on the part of the third coder. After this initial conference, the two coders did not compare their judgments until the coding was complete. Agreement for this phase of the IRR coding was 98%, $k = .75$, $p < .05$.[1] The level of agreement between the coders was thus good. All disagreements were resolved by discussion.

2.2.2. *Conceptual simulations*

There were 37 conceptual simulations throughout the protocols, an average of one conceptual simulation approximately every 9 minutes. Considering the large amount of time spent on other activities (such as choosing and setting up different visualizations,

[1] After the third coder had completed the coding, a 2x2 contingency table was constructed, counting the number of times the coders agreed there was no conceptual simulation, the number of times they agreed there was a conceptual simulation, the number of times coder 1 thought there was a conceptual simulation but coder 3 did not, and the number of times coder 3 thought there was a conceptual simulation but coder 1 did not. The nature of the data was such that there were very many instances of "no conceptual simulation," which were easy to identify (e.g., lines 2-5 of Table 7). The majority of coded utterances thus fell into the cell representing agreement on "no conceptual simulation." However, since percent agreement does not appropriately take into account agreement by chance, Cohen's kappa was used in addition to percent agreement (Cohen, 1960). Kappa of .7 is generally considered to represent satisfactory agreement.

reading off data from the visualizations, etc.), conceptual simulations occurred with sufficient frequency to be considered a real strategy used by the scientists. The frequency with which it was used compared with other strategies is discussed below.

There were 71 hypothesis statements, an average of one hypothesis approximately every 4.5 minutes. Fifty-five (77%) of these hypotheses were elaborated, i.e., the scientist further considered the hypothesis. Only elaborated hypotheses were included in subsequent analyses.

Thirty-two (86%) of the conceptual simulations occurred in reference to a hypothesis. Thus, the vast majority of conceptual simulations was coupled with the scientists' efforts to construct a satisfactory explanation of their data. We focus our analyses on how these conceptual simulations were used. (When conceptual simulation did not immediately follow a hypothesis, it was used as a problem-solving strategy, such as to resolve a difficulty in mapping between the display color and changes in velocity, to determine the circumstances under which a phenomenon might diverge from a theoretical model, or to account for a discrepancy.)

We then investigated the relative frequency of conceptual simulation compared with other strategies. Each individual utterance of data focus and tie-in with theory/domain knowledge was counted as one instance. For example, the utterance "If I look at the average of that, it's a nice clean spike" and the utterance that immediately followed it, "and I can look at the standard deviation around that and it's pretty tight right in the middle where it needs to be" were coded as two instances of data focus (average, standard deviation) because the information extracted was different in each utterance. In all other cases, the number of overall strategy uses was counted. For example, the sequence of utterances in a conceptual simulation was coded as one conceptual simulation.

First, raw frequencies for each strategy were counted, as shown in Table 9. Clearly, the most common strategy was data focus, that is, strategies that centered on the available data (as opposed to those whose focus was beyond the current data). This result is not surprising, given that the scientists' task was data analysis. However, among the strategies that focused *beyond* the immediate data, tie-in with theory/domain knowledge, conceptual simulation, and analogical reasoning/alignment occurred most frequently. We expected that expert scientists would draw on their extensive domain knowledge in understanding and analyzing data, as discussed earlier. Similarly, the use of analogical reasoning as a strategy in scientific enquiry is well documented. However, the relatively large number of conceptual simulations is striking, and provides evidence that conceptual simulation is an authentic reasoning strategy used by experts performing naturalistic tasks in their own domain.

Interestingly, proposing to collect more data and consulting colleagues occurred only rarely. Possibly, in the first case, the real-life expense (in time and money) of collecting more data made this a less attractive option than in laboratory studies of scientific reasoning, in which empirical test is frequently only a mouse-click away. Since several of the data analysis sessions involved more than one scientist, these scientists may have been less inclined to consult others, given that they were already working collaboratively (the single instance of this strategy occurred in an individual subject case).

In addition to raw frequencies, the relative frequency of each type of strategy was calculated, also shown in Table 9. For this analysis, we identified whether or not a strategy was used in reference to each hypothesis. Table 9 shows the percentage of hypotheses for which a given strategy was used *at least once* (i.e., repeated uses were not counted). As expected, the results of this analysis again show the prevalence of strategies that focus on the data. However, in terms of strategies that focus beyond the data, conceptual simulation was used as frequently as or more frequently than any other strategy. This again suggests that conceptual simulation plays a significant role in scientists' consideration and evaluation of hypotheses.

The use of analogical reasoning is also of interest. There were only two instances of general analogy, compared with 32 alignments. This result is consistent with findings of other studies in which analogy use has been found to be more "local" than "global" (Dunbar, 1997; Saner & Schunn, 1999). The use of alignment by similarity detection in relation to conceptual simulation is discussed in more detail below.

--------------------Insert Table 9 about here**--------------------**

We proposed that conceptual simulation would help scientists to resolve informational uncertainty by allowing them to evaluate their hypotheses. We suggested that upon encountering informational uncertainty, scientists would develop a possible explanation to account for it. By then running a conceptual simulation, they would be able to play out the necessary details of that explanation, creating a new representation in order to "see what happens." The resulting representation could then function as a point of comparison with the actual data representation. Insofar as the two representations match, the hypothesis would be at least supported, and therefore still offer a plausible explanation. If the relevant details do not match, the hypothesis would have to be rejected.

Trafton *et al.* (2005) have shown that scientists frequently use alignment to connect internal and external representations; consequently, we hypothesized that alignment by similarity detection would be used by these scientists to link the internal (result of the conceptual simulation) and external (phenomenon in the data) representations. Alignment would potentially allow a direct comparison between the two representations, and thus could facilitate the evaluation of the hypothesis. If this were the case, conceptual simulation would most frequently be followed by alignment (in conjunction with a hypothesis), and, to the extent that the issue is successfully resolved, alignment by similarity detection would mark the end of the reasoning chain.

The next analysis investigates this possibility by focusing on combinations of strategies. We calculated the frequencies of the transitions from one strategy to the next for all major strategies (Ericsson & Simon, 1993). In order to understand the more relevant connections between strategies, we limit our discussion to those sequences that occurred 15% or more of the time. These frequencies are represented in the transition diagram shown in Fig. 2.

The transitions of primary interest are the frequency with which conceptual simulation is followed by alignment, and the frequency with which alignment occurs at the end of the reasoning process. Here a very strong pattern is revealed. Conceptual simulations were almost always (91% of the time) immediately followed by alignment, and this sequence occurred more frequently than expected by chance, X^2 *(*4*) = 99.88, p* < .001, Bonferroni adjusted X^2s significant at *p* < .05. Alignments themselves were most likely to end the chain, a sequence that was more frequent than expected by chance, X^2 *(*4*) 15.81, p = .003*. Post-hoc comparisons showed that alignment at the end of the chain occurred significantly more frequently than alignment followed by theory, alignment, or conceptual simulation (the latter comparison was marginally significant), Bonferroni adjusted X^2s significant at *p* < .05. The difference between the frequencies of alignment followed by data focus and alignment at the end of the chain was not significant. These results suggest that the process of alignment either resolved the hypothesis under evaluation and thus terminated the chain of reasoning, or failed to resolve the hypothesis, leading the scientist to seek more information from the display.

--------------------Insert Fig. 2 about here--------------------

Several patterns emerge from the transition diagram in Fig. 2. A hypothesis was most likely to be followed by data focus, but was also followed fairly frequently by theory or directly by conceptual simulation. Data focus was almost always followed by more data focus, indicating numerous sequences in which the scientist focused explicitly on the data themselves. Theory was also most frequently followed by itself, suggesting that the scientist engaged in in-depth consideration of theoretical constructs. Theory was also a gateway to extracting information and to conceptual simulation. None of these sequences was unexpected, given the nature of the scientists' tasks. The frequency of the conceptual simulation —> alignment sequence, however, is striking, and suggests a tight coupling between the two strategies. It is in this combination of processes that the hypothesis evaluation took place.

Fig. 3 illustrates this process of conceptual simulation and alignment-based similarity detection. In this example from the Astronomy dataset, the scientists were considering the cause of some deviations from the expected pattern of velocity contours. One of them proposed a "streaming motion hypothesis"; he proposed that the existence of streaming motions might be the cause of the distortion. He then constructed a mental representation of the theoretical appearance of the velocity contours ("a perfect spider diagram"). He mentally deleted any streaming motions from this representation ("if you looked at the velocity contours without any sort of streaming motions") and identified how the lines would, then, hypothetically appear ("you'd probably expect [them] to go all the way across the ring."), i.e., he was able to "see what happened." Finally, he made a comparison between this new mental representation and the image on screen, noting that under these hypothetical circumstances, there would be no deviant segments of the contours ("without any sort of changes here in the slope"). Use of the word "here" and gestures to the screen to identify the actual deflected contour lines indicate the target of the comparison. In summary, the scientist suggested that the cause might be streaming motions, ran a conceptual simulation of the contours *without* streaming motions, noting that under these circumstances there would be no deviations in the contours, pointed out that, in contrast, there *were* kinks in the contour lines, and concluded that consequently, the streaming motion hypothesis was supported by the appearance of the data.

-------------------Insert Fig. 3 about here-------------------

2.2.3. Relationship between hypotheses and conceptual simulations

Why were only some hypotheses associated with conceptual simulation? Although almost all conceptual simulations followed a hypothesis, not all hypotheses were followed by a conceptual simulation. In this section, we attempt to tease apart why this might have been so.

In general, a hypothesis represents a scientist's best guess about an uncertain situation; however, there may be greater or lesser degrees of informational uncertainty associated with different hypotheses. If conceptual simulation is a strategy for resolving informational uncertainty, it should occur more frequently after hypotheses that relate to greater uncertainty. One way to measure the uncertainty associated with a hypothesis is to consider the scientist's knowledge about the phenomenon to which the hypothesis pertains. If there is something in the data that violates the scientist's expectations (such as a the major discrepancy between model and data), hypotheses pertaining to this phenomenon are likely to be associated with significant levels of uncertainty. If, however, the phenomenon itself is expected (e.g., in one psychology dataset the fact that subjects in the more difficult condition took longer than subjects in the control condition), hypotheses pertaining to it are likely to be associated with less uncertainty.

In order to investigate the relationship between the hypotheses and the data, the phenomenon behind each hypothesis was identified either as expected or as violating expectation. Three independent coders coded 15% of the data. Agreement between the coders was 87.5%, $k = .75$, $p < .01$; disagreements were resolved by discussion.

After the hypotheses were coded as referring to phenomena that either violated expectations or not, the use of conceptual simulation and data focus strategies to evaluate each type of hypothesis was counted. Our purpose was to determine the circumstances under which each strategy was used, consequently, only the first instance of each strategy use was counted. Table 10 shows the results of this analysis. As expected, there was no significant correlation between data focus and violate expectation ($r = .18$, $p > .1$), suggesting that data focus was a general strategy that cut across the different types of hypothesis under exploration. However, the correlation between conceptual simulation and violate expectation was significant ($r = .41$, $p < .01$). Thus, conceptual simulation appears to be a strategy that is closely associated with the investigation of hypotheses that pertain to violations of the scientists' expectations, that is, to circumstances under which there are greater levels of informational uncertainty.

--------------------Insert Table 10 about here--------------------

2.3. Summary of Study 1

The verbal protocols collected for Study 1 provided a rich dataset by which to investigate the on-line thinking of practicing expert scientists as they analyze their own data. In the course of their analysis, the scientists develop hypotheses to account for aspects of the data and then evaluate those hypotheses in light of both their theoretical knowledge and the data themselves. The analyses presented above reveal several new findings about the processes by which scientists perform this task. First, they show that scientists use conceptual simulation as a means of evaluating hypotheses and that they do so relatively frequently compared with other strategies. We propose that scientists use conceptual simulation to generate a representation of a phenomenon under hypothetical circumstances, which then serves as a source of comparison with the actual data. The comparison between this hypothetical representation and the data takes place by a process of alignment by similarity detection, which allows the scientist to evaluate whether the hypothesis under consideration remains plausible or not. Finally, these results show that the use of conceptual simulation is strongly associated with conditions of informational uncertainty, as opposed to circumstances under which the scientist's expectations were met. Study 2 investigates further the relationship between conceptual simulation and uncertainty by experimentally manipulating the scientists' expectations.

3. Study 2

Although Study 1 found a strong relationship between informational uncertainty and conceptual simulation, this relationship was correlational. Temporally, the hypotheses preceded the conceptual simulations, and conceptual simulation was more associated with phenomena that violated the scientists' expectations than phenomena that matched them. Together, these facts support our interpretation that conceptual simulation is a strategy used in situations of informational uncertainty. However, the results of Study 1 only suggest an association; they do not imply a causal relationship between informational uncertainty and conceptual simulation. In order to investigate this relationship further, we conducted a second study in which we manipulated scientists' levels of certainty about data they would be examining.

In order to retain experimental control, we conducted Study 2 as a laboratory study. However, in keeping with our goal to study the reasoning processes of practicing scientists, we replicated some of the important features of the "in vivo" Study 1. As in Study 1, our participants were expert or near-expert scientists, conducting a scientific activity in which they regularly engaged (in this case, understanding data collected by a third party). In Study 2, we focused on one domain, cognitive psychology, for which we ourselves had the necessary domain knowledge to construct realistic materials.

3.1. Method

3.1.1. *Participants*

Participants were seven cognitive psychologists (four male, three female). Three were advanced graduate students, one was a post-doctoral fellow, and three were university faculty.

3.1.2. *Tasks*

We created five tasks related to four topics within cognitive psychology—the Stroop effect, the "cocktail party effect," graph interpretation, and the effect on performance of interruptions (the interruptions topic was divided into two tasks). These topics either concerned very well-known effects or they pertained to research conducted by participants themselves or by other members of the same lab who had presented talks on this research. Thus the participants were familiar with all the topics, and were considered expert in some of them.

The format of each task was as follows: A one-page, single-spaced text described a psychological experiment—the theoretical background and rationale for the experiment (from which predictions might be drawn) and a brief method section, describing the stimuli/tasks used, the experimental conditions, the participants, and the procedure. The second page contained a bar graph representing the results of the study and a caption summarizing those results, including any relevant statistical results. An example of the tasks can be found in Appendix B.

The information in the theoretical background of the experiment was designed to lead the participant to have certain expectations about the results. There were two versions of each task, one in which the results of the experiment matched these expectations and an alternative version in which it did not. Thus two within-subjects conditions were created, an Expectation Violation (EV) condition and an Expectation Confirmation (EC) condition. The tasks were adapted from real experiments published in the psychological literature. However, they were scaled down and simplified, and in some cases the results were altered in order to create the two conditions described above.

3.1.3. *Task order*

Each participant performed one version (EV or EC) of each of the five tasks. Tasks were counterbalanced according to a Latin Square design, and the condition for each task was varied, such that each task was seen an approximately equal number of times in the EV or EC version, and each participant performed either two EV and three EC or two EC and three EV tasks. One task (the Interruptions task) was created as a sequence of two experiments; in experiment 1, the expectations were violated (EV condition), prompting a follow-up experiment, in which expectations were confirmed (EC condition). All participants performed both versions of the interruptions task.

3.1.4. *Procedure*

Participants were trained to provide talk-aloud protocols while problem solving (Ericsson & Simon, 1993). They were given the tasks one at a time by the experimenter, and they were instructed to read the materials aloud. The first page of text ended with the statement, "The results of this experiment are presented below," followed by the question participants were to answer, "What do you think could account for these results?" Thus participants were required to propose at least one hypothesis about the experimental results. The extent to which they reasoned about their hypothesis or hypotheses was left entirely to the participant. Their responses were recorded by video camera. After completing the tasks, participants were asked orally whether the results of each task were expected or unexpected to them. The protocols were transcribed and segmented, and conceptual simulations were coded, as described in Study 1.

3.2. Results and Discussion

One task, the "cocktail party effect" task, was excluded from analysis because many participants found part of the experimental manipulation and the results confusing.

3.2.1. *Inter-rater reliability*

One coder coded all of the data, and a second coder coded a subset (ten percent) of the data. (Ten percent was sufficient in this study, because of the high reliability previously established in Study 1.) Initial agreement for the conceptual simulation coding was 97%, $k = .92$, $p < .01$. Thus agreement between the two coders was extremely strong. Any disagreements were resolved by discussion.

3.1.2. *Time on task*

Participants spent an average of 49.7 minutes performing the 4 tasks, and produced an average of 422 utterances (excluding participants' initial reading of the task materials that described the study). Thus participants expended considerable time and effort performing the tasks, at least given that each task involved reasoning about only one experiment and one set of data.

3.1.3. *Use of conceptual simulation*

Overall, participants used conceptual simulation 78 times, or approximately once every 4.5 minutes, on average. This rate was approximately double that of Study 1. One possible explanation for this difference is that in Study 2, the task was *explicitly* to account for the data, whereas in Study 1, the task was to "do what you would normally do in looking at your data." Thus in Study 1, participants had to spend time determining what specific task they would perform next, how to set up the display to accomplish it, and then actually change the display. In Study 2, apart from reading the introductory text, the entire session was spent trying to explain the data.

The mean number of conceptual simulations in the EV condition was 3.8, compared with 1.9 in the EC condition. Thus, participants used conceptual simulation twice as often in the EV as in the EC condition (these were within subjects conditions). A repeated measures ANOVA on these data was significant, $F(1, 6) = 12.06, p < .05$, showing that participants were significantly more likely to use conceptual simulation when their expectations were violated than when they were confirmed. This result held across all subjects and tasks.

3.1.4. *Local EV/EC coding*

It is possible that the manipulation did not work in the predicted manner; that is, participants might not have been surprised by results in the EV condition, or might have found results in the EC condition surprising. In order to confirm that participants were indeed using conceptual simulation more frequently when their expectations were violated than when they were confirmed, a "local" EV/EC coding scheme was applied to the data. A two-stage system was used to determine whether each conceptual simulation occurred when the participant's expectations had been violated or confirmed. First, internal evidence in the protocol was used. For example, "The effect of interruption *doesn't seem too surprising,* because, um, according to theory, er the goals decay quickly" was coded as EC, whereas "That's very interesting, though, *because I would have expected something* [referring to null result]" was coded as EV. Second, if there were no *explicit* statements in a specific task's protocol that could be coded as EV or EC, the participant's self-report from the post-task interview was used. Any conceptual simulations that occurred with reference to these phenomena were coded as EC or EV accordingly, regardless of the experimental condition.

Again, one coder coded all the data, and a second, independent coder coded a subset (ten percent) of the data. Initial agreement was 98%, $k = .77$, $p < .01$, a very strong level of agreement. Any disagreements were resolved by discussion. Furthermore, for 76% of the conceptual simulations, the local coding as EV or EC matched the experimental condition. Thus, although not perfect, overall the manipulation appears to have worked as intended.

3.1.5. *Use of conceptual simulation: Local coding*

Two instances of conceptual simulation were not coded, because the participant was trying to decide whether the result was surprising or not. Sixty-eight percent of the conceptual simulations were associated with expectation violation, compared with 32% associated with expectation confirmation. A chi-square test showed that conceptual simulation was used when expectations were violated significantly more frequently than expected by chance, $X^2(1) = 12.96$, $p < .001$. This result echoes the 2:1 ratio of use produced by the experimental manipulation, and provides strong support for the hypothesis that conceptual simulation is a strategy used under conditions of expectation violation and informational uncertainty.

3.2. Summary of Study 2

Study 2 provides further evidence that scientists use conceptual simulation spontaneously when reasoning about data, and that they are more likely to do so under conditions of informational uncertainty. Whereas Study 1 provided correlational support for this hypothesis, Study 2 explicitly manipulated the participants' level of informational uncertainty, by generating situations in which either their expectations would be met or they would be violated. The results of Study 2 thus provide experimental confirmation of our interpretation of the results of Study 1.

4. General Discussion and Conclusion

These two studies show that practicing, expert scientists use conceptual simulation when working on naturalistic tasks in their own domain. This result corroborates previous research that argues for the use of mental experimentation/simulation in both historical discoveries and contemporary reasoning tasks. However, whereas historically based research depends on retrospective and narrative sources, our research finds evidence in the scientists' on-line, verbalized thinking. Furthermore, whereas other studies have identified the use of this type of reasoning by scientists of varying degrees of expertise working in domains that are not their own, and/or on artificial tasks, we have examined the behavior of professional, expert scientists working in their own domain on authentic scientific tasks.

In addition, our research demonstrates that scientists are more likely to use conceptual simulation under situations of informational uncertainty. This is shown in the “in vivo” data, where conceptual simulation was associated with the evaluation of hypotheses related to unexpected phenomena, and it is further supported in the experimental study, in which levels of informational uncertainty were explicitly manipulated. Finally, the research shows how conceptual simulation helps resolve uncertainty: Conceptual simulation facilitates reasoning about hypotheses by generating an altered representation under the purported conditions expressed in the hypothesis and providing a source of comparison with the actual data, in the process of alignment by similarity detection.

In-depth protocol studies, which use fewer participants than are generally involved in experimental research, always face questions about their generalizability. However, the consistency with which conceptual simulation was used by many individuals, as well as the range of scientific areas included in this research, suggest that the results of these two studies are likely to generalize to other scientists, at least insofar as they are performing data analysis. The use of conceptual simulation may vary in other scientific inquiry tasks, such as generating predictions from theories or designing experiments to test those theories. In general, however, we propose that scientists are likely to use conceptual simulation in situations of informational uncertainty, regardless of the specific task.

The cycle of hypothesis-conceptual simulation-alignment bears some resemblance to analogical reasoning, in that one representation (a “source”) is mapped onto another (a “target”), in order to make inferences about it. The conceptual simulation was the means by which the scientists generated the source of the comparison. The actual, displayed data representation, which the scientists were trying to understand, was the target. Alignment by similarity detection was a form of comparison that allowed the scientist to evaluate the hypothesis in order to understand something more about the underlying structure of the data representation.

There are, however, important differences between conceptual simulation and analogical reasoning. First, in the data we examined, the process of alignment was primarily based on perception, because of the visual-spatial nature of the scientists’ data; in analogical reasoning in general, however, inferences drawn about the target are not necessarily grounded in perception. Second, analogical reasoning is a memory-based strategy, i.e., similar situations that have been previously observed are recalled and used to generate predictions for a novel situation. The protocol data in these two studies, however, suggests that although the initial representation in a conceptual simulation may be grounded in memory, the transformations that are applied to it appear to be constructed afresh with each simulation. In conceptual simulation, new representations are not generated solely by reference to a familiar situation, but by taking what is known and transforming it to generate a future state of a system. Thus conceptual simulation may be considered a form of model *construction*, which is likely to occur when no easily accessible, existing source for analogy is available. This situation may be similar to that identified by Griffith and colleagues, who propose that when model search and analogy fail, scientists construct and manipulate mental models (e.g., by means of general structural transformation) (Griffith, Nersessian, & Goel, 2000).

Like analogical reasoning, conceptual simulation can also be considered a type of reasoning with inductive mental models (e.g., Nersessian, 1992b; Schwartz & Black, 1996b). Although the term "mental model" is used frequently, there is wide-scale disagreement about precisely what constitutes a mental model. In our view, mental models are dynamic and "runnable." This means that the components of the model can be set in motion and their behavior and changes of state can be observed, in a process that mirrors observations of the physical components of a tangible model. The output of running a mental model is an *inference* about the outcome of a particular converging set of circumstances. By animating their mental models, people are able to simulate a system's behavior in their "mind's eye" and to predict one or more possible outcomes, even for situations in which they have no previous experience (Gentner, 2002). Conceptual simulation involves transforming ("running") a representation, and inspecting the output, a changed representation that becomes the basis for inferences about the data.

Conceptual simulations, like other kinds of mental model, rely on *qualitative* relationships, such as signs and ordinal relationships, relative positions and so on, rather than precise numerical representation. In general, mental models are particularly instrumental in guiding problem-solving when people lack a formal scientific understanding of a domain (e.g., Forbus, 1983; Gentner & Gentner, 1983; Kieras & Bovair, 1984). Although the expert scientists in our studies did not lack formal scientific understanding, they did lack the precise knowledge to immediately solve the informational uncertainty they were experiencing. Conceptual simulation seems to have allowed them to engage in causal reasoning about a system, even in the midst of this informational uncertainty.

As a form of "what if" reasoning, conceptual simulation is also strongly related to the type of thought experiment discussed by Nersessian (1992b). Nersessian also interprets thought experiments as a form of reasoning with mental models and proposes that such mental models are "temporary structures constructed in working memory for a specific reasoning task." We have argued that conceptual simulations are similarly constructed to meet a specific, temporary need. Nersessian argues for the importance of this type of reasoning in instances of major conceptual change in scientific discovery. Unlike these thought experiments, which may lead to large-scale conceptual change, conceptual simulations may be considered small-scale, or "local," thought experiments. Although we did not observe any major conceptual change in our data, we did witness numerous instances of scientists using conceptual simulation to get "unstuck" when they had reached an impasse in understanding their data, and in this sense, conceptual simulation may serve a similar function of helping a scientist move beyond what is currently known.

In general, experts' domain knowledge provides them with many existing solutions and analogs upon which to draw during problem-solving (e.g., Chi, Feltovich, & Glaser, 1981). Yet we found true experts generating conceptual simulations, rather than retrieving solutions from memory. We propose that conceptual simulation will be used by experts when they are working either outside their immediate area of expertise or on their own cutting edge research—that is, in situations that go beyond the limits of their current knowledge. This interpretation meshes with Schraagen's observation that conceptual simulation was used on a task in the domain of gustatory psychology by psychologists expert in domains *other* than gustatory psychology, but not by novices or by experts *within* the gustatory domain (Schraagen, 1993). Although Schraagen was led to conclude that it is therefore an intermediate strategy, his results are not inconsistent with our suggestion that experts working on a truly novel task in their own domain would engage in conceptual simulation. The extent to which novices are able to productively use conceptual simulation in situations of uncertainty remains a matter for investigation. We predict, however, that novices will be less capable of generating conceptual simulations because they lack domain knowledge, and that therefore they will use fewer conceptual simulations than experts.

There are very few studies of expert scientists performing "real" scientific tasks. In his pioneering "in vivo" study of molecular biologists, Dunbar asked, "How do scientists really reason?" (Dunbar, 1995). Our studies contribute further to our understanding of how scientists really reason. Frequently, studies of experts employ problems that are well-understood for an expert and that can be solved by recalling either this very problem (i.e., by model-based search) or another that shares the same deep structure (i.e., by analogy, cf. Chi et al., 1981). In contrast, our studies show experts reasoning about problems for which neither they nor anyone else knows the answer. In such circumstances, they must construct new models "on the fly," tailor-made to the problem and its context. This strategy of conceptual simulation is similar to mental model-based strategies used by laypeople in reasoning about the everyday world. Expert scientists, however, have the domain knowledge that allows them to generate predictions that are accurate and therefore useful in the context of scientific problem solving.

With the current emphasis in science education reform on authentic practice (NRC, 1996), these studies have practical implications for efforts to improve science in the classroom. Not only does current educational theory suggest that instruction should be situated in the context of authentic scientific questions to which students genuinely desire to learn the answer (Barron, Schwartz, Vye, Moore, Petrosino, Zech, & Bransford, 1998), but also that students be encouraged to use the tools and strategies of real scientific practice. Research has already shown the value of having students generate predictions prior to conducting experiments (White, 1993); however, the prediction generation process itself has been largely unexplored. It is possible that qualitative reasoning strategies, such as the use of mental models and conceptual simulation, can be explicitly taught to students, providing them with a more formal means to generate predictions, specify their implications, evaluate their accuracy, and identify potential causes of discrepancies.

There have been many myths about how scientists operate, including the idea of the “lone scientist” toiling in isolation, the belief that scientific discovery is the result of genius, inspiration, and sudden insight, the assumption that hypotheses should always precede experimentation and observation, and especially the notion that scientists are unbiased processors of objective data. Research in cognitive science has helped to dispel many of these myths; the current study contributes further to our understanding of the processes by which scientific knowledge actually develops in the real world. It provides evidence to support the claim that science advances not through the use of mysterious and inexplicable processes unique to a particular group of geniuses but through the systematic use of everyday processes. Conceptual simulation—a specific type of qualitative mental model—is one such everyday reasoning process.

References

Azmitia, M., & Crowley, K. (2001). The rhythms of scientific thinking: A study of collaboration in an earthquake microworld. In K. Crowley, C. D. Schunn & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 51-82). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Barron, B. J. S., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., & Bransford, J. D. (1998). Doing with understanding: Lessons from research on problem-and project-based learning. *Journal of the Learning Sciences, 7*(3 & 4), 271-213.

Brown, J. R. (2002). Thought experiments. *The Stanford Encyclopedia of Philosophy (Summer 2002 Edition)*. Available at: http://plato.stanford.edu/archives/sum2002/entries/thought-experiment/. Accessed May 24, 2006.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chinn, C., & Malhhotra. (2001). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn & T. Okada (Eds.), *Designing for science: Implications from everyda,y classroom, and professional settings*. Mahwah, NJ: Erlbaum.

Clement, J. (2002a). *Protocol evidence on thought experiments used by experts*. Paper presented at the 24th Annual Conference of the Cognitive Science Society, Fairfax, VA.

Clement, J. (2002b). Step-wise evolution of mental models of electric circuits: A "learning-aloud" case study. *The Journal of the Learning Sciences, 11*(4), 389-452.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measurement, 20,* 37-46.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 365-395). Cambridge, MA: MIT Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward & S. M. Smith (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461-493). Washington, DC, USA: American Psychological Association.

Einstein, A. (1979). *Autobiographical notes*. Peru, IL: Open Court Publishing Company.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), 725-747.

Ericsson, K. A., Krampe, R. T., & Tesch-Roemer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. (2nd ed.). Cambridge, MA: MIT Press.

Forbus, K. (1983). Reasoning about space and motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 53-74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Forbus, K. (2002). Qualitative modeling of common sense understanding. *Cognitive Science Society Virtual Colloquium Series.*

Forbus, K., & Gentner, D. (1997). *Qualitative mental models: Simulations or memories?* Paper presented at the Eleventh International Workshop on Qualitative Reasoning, Cortona, Italy.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou and A. Ortony (Ed.), *Similarity and analogical reasoning*. New York, NY: Cambridge University Press.

Gentner, D. (2002). Analogy in scientific discovery: The case of Johannes Kepler. In L. Magnani & N. J. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp. 21-39). New York: Kluwer Academic/Plenum Publisher.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens, (Eds.), *Mental models* (pp. 99-129). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45-56.

Griffith, T. W., Nersessian, N. J., & Goel, A. (2000). *Function-follows-form transformations in scientific problem solving*. Paper presented at the The 22nd Annual Conference of the Cognitive Science Society, Philadelphia, PA.

Hayes, P. J. (1988). Naive physics 1: Ontology for liquids. In A. M. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence* (pp. 251-269). San Mateo, CA: Morgan Kaufmann, Inc.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation (vol. 19)* (pp. 59-87). New York: Academic Press.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*(3), 295-395.

Ippolito, M. F., & Tweney, R. D. (1995). The inception of insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 433-462). Cambridge, MA, USA: MIT Press.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning how to operate a device. *Cognitive Science, 8*, 255-273.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Klahr, D., Dunbar, K., & Fay, L. (1990). Designing good experiments to test bad hypotheses. In J. Shrager & P. Langley (Eds.), *Computational models of discovery and theory formation* (pp. 355-402). San Mateo, CA: Morgan-Kaufman.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125*, 524-543.

Kulkarni, D., & Simon, H. A. (1988). The process of scientific discovery: The strategy of experimentation. *Cognitive Science, 12*, 139-176.

Lozano, S. C., & Tversky, B. (2006). Communicative gestures facilitate problem solving for both communicators and recipients. *Journal of Memory and Language, 55*(1), 47-63.

Nersessian, N. J. (1992a). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Cognitive models of science* (pp. 3-44). Minneapolis, MN: University of Minneapolis Press.

Nersessian, N. J. (1992b). In the theoretician's laboratory: Thought experimenting as mental modeling. *PSA, 2,* 291-301.

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J. Nersessian & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 5 - 22). New York: Kluwer Academic/Plenum Publishers.

NRC. (1996). *National science education standards.* Washington, DC: National Research Council.

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21*(2), 109-146.

Popper, K. R. (1956). *The logic of scientific discovery (Rev. Ed)*. New York: Basic Books.

Qin, Y., & Simon, H. A. (1990). *Imagery and problem solving*. Paper presented at the Twelfth Annual Meeting of the Cognitive Science Society.

Saner, L., & Schunn, C. D. (1999). *Analogies out of the blue: When history seems to retell itself.* Paper presented at the 21st Annual Conference of the Cognitive Science Society.

Schraagen, J. M. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17(2)*, 285-309.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23*(3), 337-370.

Schunn, C. D., Kirschenbaum, S. S., & Trafton, J. G. (under review). The ecology of uncertainty: Sources, indicators, and strategies for informational uncertainty.

Schwartz, D. L., & Black, J. B. (1996a). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology, 32*(2), 154-219.

Schwartz, D. L., & Black, J. B. (1996b). Shuttling bewteen depictive models and abstract rules: Induction and fallback. *Cognitive Science, 20,* 457-497.

Shepard, R. (1988). The imagination of the scientist. In K. Egan & D. Nadaner (Eds.), *Imagination and education* (pp. 153-185). New York and London: Teachers College Press.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (2005). Connecting internal and external representations: Spatial transformations of scientific visualizations. *Foundations of Science, 10*, 89-106.

Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L., Schunn, C. D., & Kirschenbaum, S. S. (2006). The relationship between spatial transformations and iconic gestures. *Spatial cogntion and computation, 6*(1), 1-29.

Trickett, S. B., Trafton, J. G., Saner, L., & Schunn, C. D. (In press) *"I don't know what's going on there":* The use of spatial transformations to deal with and resolve uncertainty in complex visualizations.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*(1), 75-100.

White, B. Y. (1993). Thinkertools: Causal models, conceptual change, and science education. *Cognition & Instruction, 10*(1), 100 p.

Williams, B. C., & de Kleer, J. (1991). Qualitative reasoning about physical systems: A return to roots. *Artificial Intelligence, 51*, 1-9.

Acknowledgments

Work on this project was supported by ONR grants M12439 and N0001403WX30001 to the second author. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Navy. This research was the first author's dissertation. She would like to thank all the members of her dissertation committee: Chris Schunn, Debbie Boehm-Davis, and Brenda Bannan-Ritland for their comments and guidance. Thanks also to Erik Altmann, Audrey Lipps, and Bill Lile for comments on earlier drafts of this work, to Peter Cheng, Nancy Nersessian, Dan Schwartz, and an anonymous reviewer for their reviews of this manuscript, and to Raj Ratwani for assistance with coding.

Appendix A
Conceptual Simulation Training

We want you to read through every line in the protocol and mark it in the following way. First, you need to ask whether the speaker is creating a new mental representation. One way to think about this is to determine whether he or she is referring directly to what is currently on display on the computer screen. If so, there is no new mental representation. If the scientist is referring to something in his or her head, you should note that as a new representation. The new representation could refer to a memory of something he or she has already seen, or it could refer to a theoretical construct, or it could refer to a hypothetical situation that the scientist is constructing for the first time.

When you identify a new representation, you should code the utterances that follow it, using the spatial transformation coding scheme. That is, if the scientist mentally manipulates or transforms the starting representation spatially, you should code that utterance accordingly. Finally, immediately after any utterances that you have coded as spatially transforming the starting representation, you should examine the next utterance(s) to determine whether there is a "result" of the transformations, or an ending representation that is different from the starting representation. If you find all three components of this sequence, you should code each utterance as conceptual simulation (CS). For any utterance that is not a part of this type of sequence, you should code it as no conceptual simulation (No CS).

Here are two examples from the astronomy dataset that illustrate this coding scheme. In the first example, note that although the scientists are trying to explain a particular phenomenon by proposing different hypothetical situations, and although a new representation is generated, there is no conceptual simulation, because no spatial transformations are applied to the new representation. The entire sequence (refer to new representation—refer to mentally transforming representation—refer to result of representation) is not present. In the second example, there are a reference to a new representation, reference to several spatial transformations performed on that representation, and reference to an end result of those transformations. Consequently, each of those utterances is coded as CS.

Example 1:

Utterance (scientist 1)	Utterance (scientist 2)	CS Coding	Explanation (training purposes only)
That might just be gas blowing from the star-forming regions		No CS	Scientist is trying to explain what might account for "stuff all over here" identified previously
	But that's not a star-forming region, though, at the centre left	No CS	Identifies feature of current display
Centre left		No CS	Searches display to identify area of interest
	That one	No CS	Identifies area of interest
Maybe this stuff is just sort of infalling		No CS	Spatial transformation: mentally moves "stuff" from one location to another. However, coded as *no CS* because it does not follow a reference to a new representation, or lead to a changed representation
I mean, you know, if there's a big gas cloud…		No CS	New representation. Coded as No CS because the representation is not transformed.
	Infalling as a big blob?	No CS	Queries explanation
Why not? Why not? Why can't gas infall as a big blob?		No CS	Reiterates explanation
	The pressure thing tends to push them apart, though	No CS	States domain knowledge

	I mean, it seems like there should be a kinematic reason for that	No CS	States domain knowledge
	Ah, I don’t see what it is	No CS	Unable to resolve

Example 2:

Utterance (scientist 1)	**Utterance (scientist 2)**	**CS Coding**	**Explanation (training purposes only)**
	It seems like the H1 disk here is offset	No CS	Scientist is looking at image of galaxy and interpreting it
The H1 disk is offset…Can you have that happen?		No CS	Questions interpretation
	Sure, I, I, well, I think you can actually	No CS	
	Umm, I mean, remember, these things are in the elliptical orbits	CS	New representation (displayed image does not show anything about orbits)
	Things may be falling kind of inward as they’re going around the orbits,	CS	Spatial transformation: mentally moves matter from one place to another, and moves it around in orbit
	The gas pressure is sort of driving the H1 out a little bit more	CS	Spatial transformation: mentally moves the H1 from one location to another
	And when it falls back in because of the dissipation going on	CS	Spatial transformation: Mentally moves H1 from one location to another
	You could have it offset that way	CS	End result: offset disk

Appendix B
Sample Materials for Study 2

Interruptions

Altmann & Trafton (in press) have suggested that there are 3 things that memory for goals depends on:

1) Rehearsal (you may need to rehearse your goal to remember it later)
2) Cues in the environment (i.e., something in the environment may remind you what your goal was)
3) The fact that individual goals decay quite quickly (in seconds)

Recently, Trafton ran an experiment to examine how rehearsal affected resuming a task after an interruption. The task was set up so that participants were working on a goal as they got interrupted. The experiment used two tasks, a primary task that participants worked on most of the time and a secondary task that was the "interrupting" task. The primary task was a complex resource allocation task that had many different goals and many different things participants could do at any point in time. The secondary task was a dynamic categorization task (the Ballas task, a lot like Argus).

Participants worked on the primary task for approximately 20 minutes. There were 10 interruptions throughout the 20 minute scenario. Each interruption followed a mouse-click to ensure that a participant was working on a goal (or, rather, to ensure the participant was actively working on some task, not just thinking or spacing out). There were two conditions:

- A No Warning condition (NW) where participants were immediately taken to the secondary task.
- A Warning condition (W) where participants were given 8 seconds to "prepare" for the secondary task. Participants were warned they were switching to the secondary task by a set of "eyeballs" that appeared on the screen. Once the eyeballs showed up, participants were not able to work on the primary task and were told to "remember what they were working on."

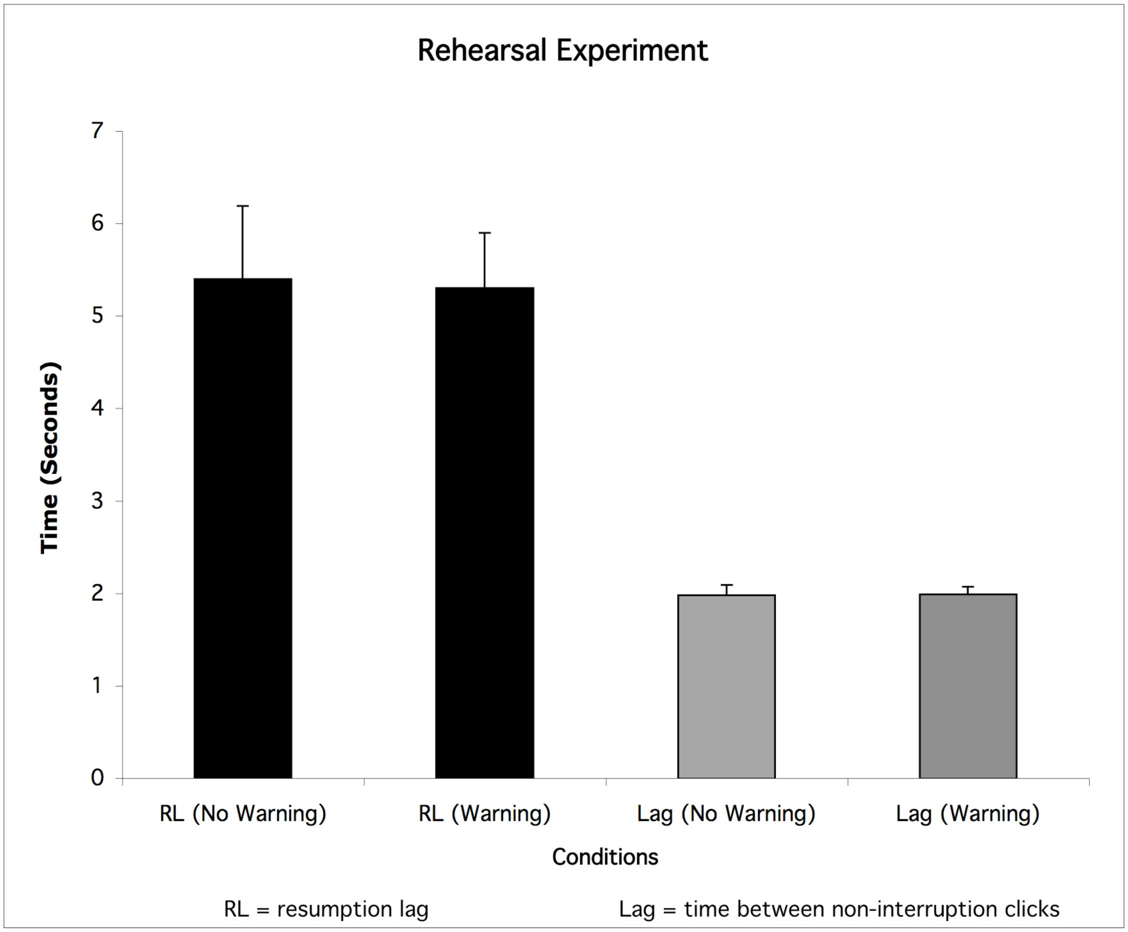
All participants were told that when they came back to the primary task, they were to resume where they left off (i.e., to remember the goal they were working on).

There were 10 subjects in each condition.

The secondary task lasted approximately 45 seconds.

According to Altmann & Trafton, the Warning condition was expected to have a much faster resumption lag (RL) than the No Warning condition. (A resumption lag is the time it takes people to resume a task after being interrupted; a regular lag is the time between key strokes without an interruption).

The results of this experiment are presented below. What do you think could account for these results?



Error bars are standard error of the mean.

There is a highly significant effect of interruption: resuming a task after an interruption takes much more time than lags measured during the primary task.

There is no effect of condition (F < 1).

Fig. 1. Screen snapshot of computational fluid dynamics data

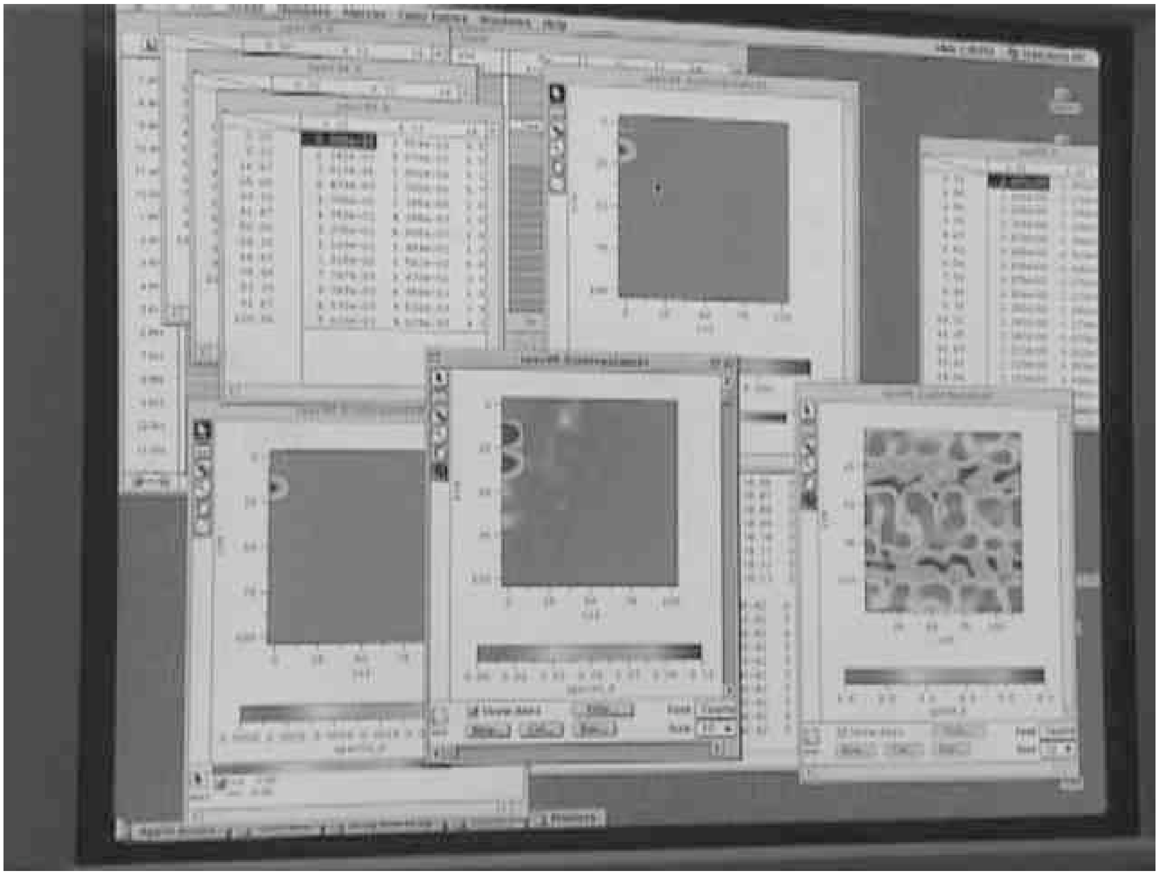


Fig. 2: Transition diagram showing the relationships among strategies. Percentages show the frequency with which one strategy followed another.

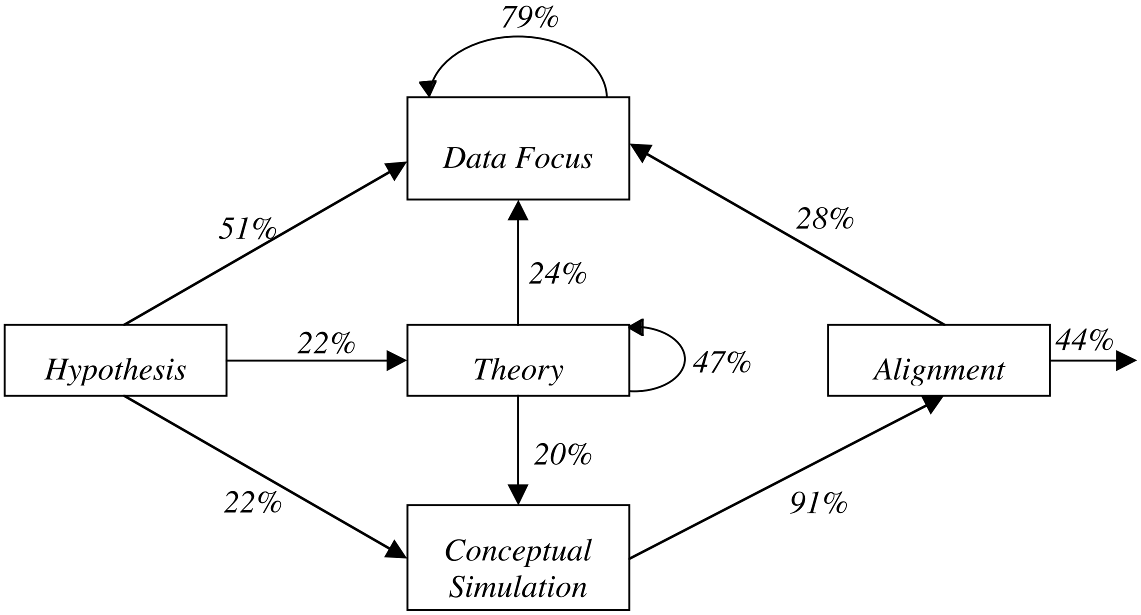
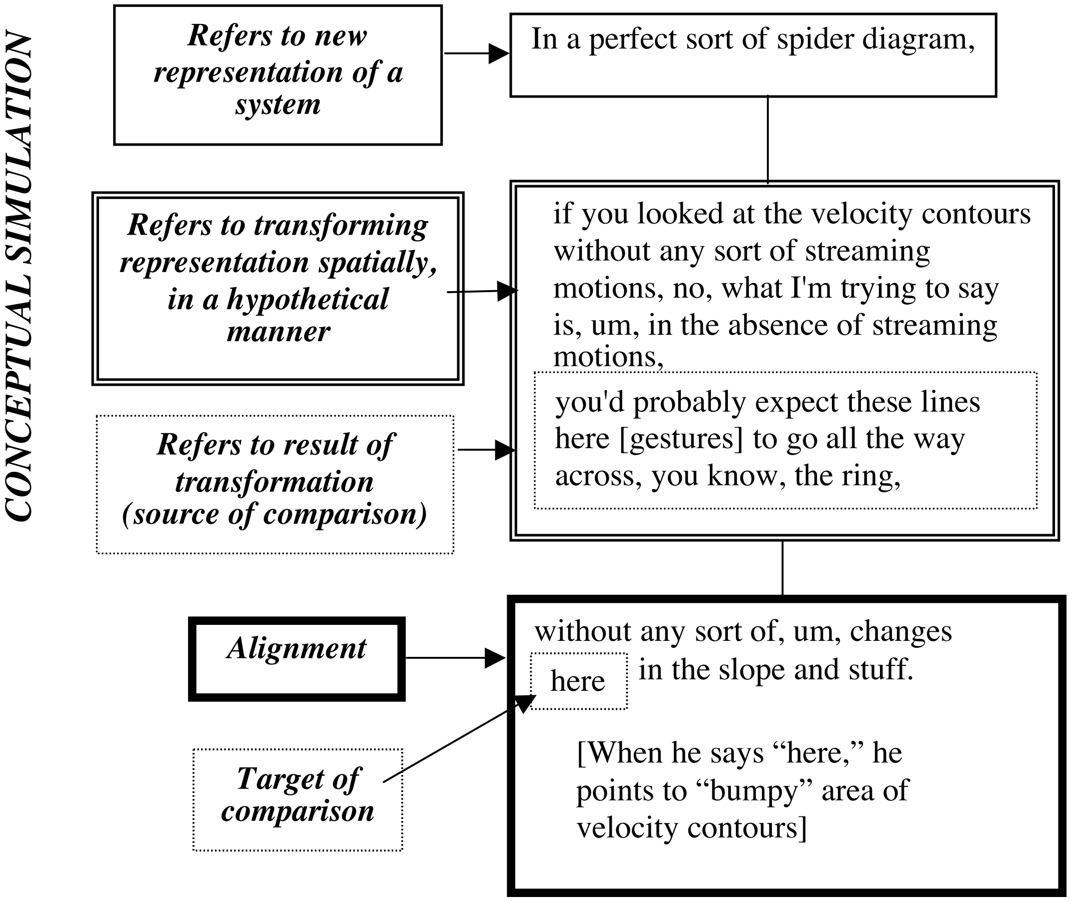


Fig. 3: Conceptual simulation used as source of comparison in alignment process. An anomaly in the external display functions as the target of the comparison, and the scientist uses conceptual simulation to generate the source of the comparison.


CONCEPTUAL SIMULATION
Refers to new representation of a system
Refers to transforming representation spatially, in a hypothetical manner
Refers to result of transformation (source of comparison)
Alignment
Target of comparison

In a perfect sort of spider diagram,

if you looked at the velocity contours without any sort of streaming motions, no, what I'm trying to say is, um, in the absence of streaming motions,

you'd probably expect these lines here [gestures] to go all the way across, you know, the ring,

without any sort of, um, changes here in the slope and stuff.

[When he says "here," he points to "bumpy" area of velocity contours]

Table 1: Dataset characteristics

Dataset	On-task Utterances	% of Total Utterances	Number of Scientists	Total Relevant Time
Astronomy	656	76	2	49 minutes
CFD submarine	437	42	1	39 minutes
CFD laser 1	172	43	1	15 minutes
CFD laser 2	184	74	1	13 minutes
fMRI	215	72	2	55 minutes
Neural spikes	217	64	2	54 minutes
Psychology 1	481	89	3	31 minutes
Psychology 2	916	64	2	75 minutes

Table 2: Characteristics of individual data analysis sessions

Domain	Research Stage	Data Type	Data	Data Source	Task Description
Astronomy	exploratory	visual	velocity contour lines laid over optical data	telescope observations	Understand flow of gas in galaxy
CFD submarine	confirmatory	visual	2D line plots	computational model	Understand model in relation to empirical data collected by a different researcher
CFD laser 1	confirmatory	visual	contour plots or Fourier decompos-ition	computational model	Understand growth rate and sequence of different modes
CFD laser 2	confirmatory	visual	contour plots or Fourier decompos-ition	computational model	Follow-up Laser 1
fMRI	confirmatory	visual	structural or functional brain images	controlled experiment	Identify areas of neural activity; evaluate experiment predictions
Neural spikes	exploratory	visual	neural spikes	surgical observations	Isolate single cell firings in order to distinguish real from spurious neurons
Psychology 1	exploratory	numeric	numerical in spreadsheet	controlled experiment	Seek evidence for strategies among subjects
Psychology	exploratory	numeric	numerical	controlled	Understand

2			in spreadsheet	experiment	relationship between subject and model data

Table 3: Examples of conceptual simulation (CS)

Source	Utterances	Code	Explanation
Astronomy	*In a perfect sort of spider diagram*	CS	Reference to new representation (“spider diagram”)
	if you looked at the velocity contours without any sort of streaming motions, no, what I’m trying to say is, um, in the absence of streaming motions	CS continued	Reference to transforming representation (mentally removing existing streaming motions)
	you'd probably expect these lines here [gestures] *to go all the way across, you know, the ring*	CS continued	Reference to result (sees what happens)
CFD submarine	*It is conceivably possible that this curve is floating around all over the place, and what they’re showing is an average* [scientist is looking at a graphical representation (a curve) that represents the turbulence]	CS	Reference to new representation (“this curve”)
	so if this thing is really floating around that much, just up and down, and I’m at the extreme end, and if I average all of this stuff,	CS continued	Reference to transforming representation
	then I may actually still get the curve right	CS continued	Reference to result (sees what happens)

Table 5: Examples of data focus strategies

Source	Utterance	Explanation
fMRI	*We can find out what the z-score of that one is, too. Let's see, it's 4.22, 4.23*	Read off data
Astronomy	*Actually, I know that the, this is a naturally weighted method. If we look at the robust, let's look at the robust weighted method*	Change visualization
Psychology 1	*So I mean this is a post-hoc hypothesis, that we could verify by looking at the patterns*	Examine additional available data
Psychology 2	*We have an outlier there. We can get rid of that guy probably….That's more than three times the mean standard deviation*	Tweak data (remove outlier)

Table 6: Examples of empirical test strategies

Source	Utterance	Explanation
Astronomy	*Do you think it's worth getting some more [telescope] time, just to do an offset plane, or offset velocity?*	Collect more (observational) data
fMRI	But we also have to be cognizant of the limitations of the equipment we're working with. *And we are, like I said, when we collect data again, for instance, we are going to get the whole brain.*	Collect more (experimental) data
CFD (submarine)	*That means I have to tweak an input parameter on the flow code. And then re-run it [the model].*	Run computational model

Table 7: Examples of analogy and alignment (relevant phrases that pinpoint the actual analogy or alignment are in italics; utterances in Roman type are for context only, and were not coded as analogy/alignment)

Source	Utterance	Explanation
Astronomy	*Think of this* [points to part of ring galaxy] *as a spiral arm*	Explicit analogy between "spiral arm" (source) and "this" (ring galaxy); scientist is using the concept of a spiral arm to make inferences about the behavior of a system that is *not* a spiral arm
CFD (laser 2)	So [0-2] is going to be way below the black line…but he's gonna grow at roughly the same rate [as 2-0] *which is what you would expect*	Alignment: scientist aligns growth rates of one mode (0-2) with another (2-0), and with theoretical expectations
CFD (laser 2)	The high modes are supposed to take off. They're supposed to run faster, which means that if that guy took off first, then he should be like, dominating the whole action. Now the only possible way that that can't happen is if this guy has some source somewhere, that he's, like, being fed. *And he is being fed…by the difference of these two guys.*	Alignment: scientist aligns his expectation that mode must be being "fed" with the data representation, which indicates that the mode *is*,in fact, being fed.
CFD (submarine)	You know what, this is an experiment that sets in a, in a tube, and they've got struts holding that sucker up onto the floor. I wonder if I'm seeing the wake of the struts, which, of course, we don't have on our computational model—so that's why we don't see a dip. *But we're still off by a good few percent, way off there…*	Alignment: scientist aligns the experimental data with his image of the model data, after accounting for the presence of the struts; the alignment shows there are still significant differences between the model and experimental data
Astronomy	It's, I mean, it seems to make sense, if that's operating, if it's all the	Alignment: scientist aligns the output of his chain of reasoning

	same velocity, it's probably more or less a rigid body, so that the whole thing is—I mean, so does that make sense? *No, it doesn't really, nah, it's not necessarily a right body…*	that suggests a rigid body with the actual data, which does not show a rigid body

Table 8: Illustration of initial approach to coding conceptual simulations (CS) (source: Laser 2)

	Utterance	Coding
1	Was outrun by the next one down	
2	And I don't know	
3	I just don't know	
4	I'll haveta get someone else's interpretation of that	
5	I don't understand that	
6	The high modes take off	CS: New mental representation of beginning state (display shows end state)
7	They're supposed to run faster	CS: Describes new representation
8	Which means if that guy [mode 1]took off first	CS: Mentally follows growth path of mode 1
9	Then he should be like dominating the whole action	CS: Mentally places mode 1 in relation to mode 2
10	Now the only possible way that that *can't* happen	CS: Mentally undoes growth path of mode 1
11	Is if this guy [mode 2] has some source somewhere	CS: Mentally adds source to representation of mode 2
12	That he's like, being fed	CS: Mentally adds source to representation to mode 2
13	And he is being fed	Alignment
14	The only way he gets fed is by the difference of these two guys [additional modes]	Alignment
15	OK, the, the physics of this is	
16	The physics of this is any two modes that can add up	
17	Because of their non-linear action	
18	Feed the next one	
19	So the mode interacts with itself	
20	One-one, to produce a two	
21	But one and two can interact and produce a three	
22	But three, ah, three minus two can also produce one	

23	So they sort of interact among themselves	

Table 9: Frequencies of occurrence of hypothesis-evaluation strategies—total number of uses (raw frequency) and percentage of all hypotheses for which strategy was used (relative frequency). Note that because more than one strategy might be used with a given hypothesis, these percentages sum to more than 100

Strategy	Raw Frequency	Relative Frequency (% hypotheses)
Data focus	229	65%
Tie-in with theory	51	35%
Alignment	32	47%
Conceptual simulation	32	46%
Empirical test	3	.05%
“Far” analogy	2	.04%
Consult Colleague	1	.02%

Table 10: Percentages of violate expectation and no discrepancy hypotheses for which conceptual simulation and data focus were used

	Violate Expectation	No Discrepancy
Conceptual Simulation	64%	21%
Data focus	61%	79%